INFORMATION AND CONTROL 19, 216–223 (1971)

A Few Remarks on the Index of Context-free Grammars and Languages

J. GRUSKA*

Department of Computer, Information, and Control Sciences
University of Minnesota

A hierarchy of context-free grammars and languages with respect to the index of context-free grammars is established and the undecidability of the basic problems is proven.

1. NOTATION

Let $G = \langle V, \Sigma, P, \sigma \rangle$ be a context-free grammar (in short, a grammar) where $\Sigma \subset V$ is the set of terminals, $V - \Sigma$ the set of nonterminals $\sigma \in V - \Sigma$ is the initial symbol of G and $P \subset (V - \Sigma) \times V^*$ is a finite set of productions of G. $L(G) = \{x; \sigma \overset{*}{\Rightarrow} x \in \Sigma^*\}$ is the context-free language (in short, a language) generated by G.

Let ϵ denote the empty word, $| x |$ the lenght of word x, and I the set of positive integers.

Following Brainerd (1968), the index of a derivation τ, in short Ind(τ), where

$$\tau : w_1 , w_2 ,..., w_k ,$$

is the smallest integer i_0 such that neither of the words w_i, $1 \leqslant i \leqslant k$, has more than i_0 occurrences of nonterminals. For an $x \in L(G)$, $\text{Ind}(x) = \min\{\text{Ind}(\tau); \tau$ is a derivation of x from σ in $G\}$.

For a grammar G and a language L let $\text{Ind}(G) = \max\{\text{Ind}(x); x \in L(G)\}$; $\text{Ind}(L) = \min\{\text{Ind}(G); L(G) = L\}$. If $\text{Ind}(G) < \infty$ ($\text{Ind}(L) < \infty$), then the grammar G (the language L) is said to be of finite index; otherwise of infinite index.

* Present Address: Mathematical Institute, Slovak Academy of Sciences, Bratislava, Czechoslovakia.

2. Salomaa's Problem

In a recent paper, Salomaa (1969) raised the question of whether there are two grammars which generate the same language but only one of them is of finite index. The answer is in the positive. Indeed, by Salomaa (1969), there exists a grammar G_1 with the infinite index which generates the Dyck language L_0 over the alphabet $\{0, 1\}$. L_0 is a deterministic language and therefore a grammar G_2 generating $\{0, 1\}^* - L_0$ exists. Combining these two grammars we get a grammar with the infinite index which generates the language $\{0, 1\}^*$ having the index 1.

3. Finite Index Languages

By Salomaa (1969), there is a context-free language of infinite index. Languages of finite index form a very natural class of languages which has been studied in several papers under different names (superlinear languages, derivation-bounded languages, semilinear languages and so on).

Theorem 1. *The class of finite index languages is the small full* AFL *which contains linear languages and is closed under substitution.*

Proof. Each finite index language is a derivation-bounded language and by Ginsburg and Spanier (1968) the class of derivation-bounded languages forms a full AFL mentioned in the Theorem. On the other hand, each derivation-bounded language can be obtained from linear languages by substitution and, therefore, is of a finite index.

Several infinite hierarchies of finite index languages have recently appeared in the literature (Greibach, 1969; Gruska, 1969). In the following theorem, a new hierarchy depending on the index of languages is proven.

Theorem 2. *For every* $n \in I \cup \{\infty\}$ *there is a language* L_n *such that* $\mathrm{Ind}(L_n) = n$.

Proof. Let L_0 be the Dyck language over the alphabet $\{0, 1\}$; i.e., the language generated by the grammar

$$\sigma \to 0\sigma 1, \qquad \sigma \to \sigma\sigma, \qquad \sigma \to \epsilon.$$

By Salomaa (1969), $\mathrm{Ind}(L_0) = \infty$. Trivially, $\mathrm{Ind}(\{\epsilon\}) = 1$. For n finite and $n > 1$, let $L_n = L_0 \cap (0^*1^*)^{2^{n-1}}$. The theorem will be proved by showing that

$\mathrm{Ind}(L_n) = n$. Since L_n is generated by the grammar G_n with the initial symbol σ_1 and the rules[1]

$$\sigma_i \to 0\sigma_i 1 \mid \sigma_{i+1}\sigma_{i+1} \mid \epsilon, \qquad 1 \leqslant i \leqslant n-1,$$
$$\sigma_n \to 0\sigma_n 1 \mid \epsilon,$$

we have immediately $\mathrm{Ind}(L_n) \leqslant n$. To complete the proof, it remains to show that $\mathrm{Ind}(L_n) \geqslant n$.

To that end, let $G = \langle V, \{0, 1\}, P, \sigma\rangle$ be a grammar such that $\mathrm{Ind}(G) = \mathrm{Ind}(L_n)$, $L(G) = L_n$ and, moreover, all rules of G have either the form $A \to uBv$ or $A \to uBCv$ with u, v being terminal words and B, C being nonterminals. Clearly, such a G does exist. We can also assume that G is a reduced grammar, i.e., all nonterminals are reachable and generate some terminal words.

Let $m = \max\{|\alpha|; A \to \alpha \text{ is in } P\}$ and let n_0 be the number of nonterminals of G.

Let $\psi : \{0, 1\}^* \to \{0, 1\}^*$ be the mapping defined by $\psi(x) = \psi(y_1 y_2)$, if $x = y_1 01 y_2$ and $\psi(x) = x$ otherwise.

If $x \in L_n$ and $x = yz$, then $\psi(y) \in \{0\}^*$ and $\psi(z) \in \{1\}^*$. From that and from the structure of words in L_n, it follows

if, in G, $A \overset{*}{\Rightarrow} xAy$ for a nonterminal A and terminal words x, y, then $\psi(x) = 0^k$, $\psi(y) = 1^k$ for some $k \geqslant 0$. (1)

If in a derivation tree of a word x no path contains a nonterminal twice, then $|x| \leqslant m^{n_0}$. In view of (1) this in turn implies

$$\text{if} \quad A \overset{*}{\Rightarrow} x \text{ in } G, \qquad \text{then} \quad |\psi(x)| \leqslant m^{n_0}, \tag{2}$$

and, as a corollary,

$$\text{if} \quad A \overset{*}{\Rightarrow} x \in \{1\}^*\{0\}^*, \qquad \text{then} \quad |x| \leqslant m^{n_0}. \tag{3}$$

Now let $N > 3m^{n_0+1}$ be an integer and let y_i, $i \geqslant 0$, be words defined by

$$y_0 = \epsilon, \qquad y_{i+1} = 0^{2N} y_i 1^N 0^N y_i 1^{2N}, \qquad i \geqslant 0.$$

For every A in V, let $\nu(A) = \{A\} \cup \{x; A \overset{*}{\Rightarrow} x \in \{1\}^*\{0\}^*\}$. By (3), $\nu(A)$ is a finite set.

[1] The productions are written in an abbreviated form $A \to \alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_n \mid$ instead of $A \to \alpha_1$, $A \to \alpha_2, \ldots, A \to \alpha_n$.

Let us now construct a new grammar G' from G by replacing every production $A \to \alpha$ of G by the set of productions $\{A \to z; z \in \nu(\alpha)$, where $\nu(xa) = \nu(x)\nu(a)$, for any word x and symbol a.

Clearly, $L(G') = L_n$, $\mathrm{Ind}(G') = \mathrm{Ind}(L_n)$. The length of right sides of productions of G' is not more than m^{n_0+1} and (1)–(3) hold also for G'. For the rest of this proof we will deal only with the grammar G' and, therefore, all derivation concepts, for example $\overset{*}{\Rightarrow}$, refer to G'.

Assume now for a moment that the following lemma has already been proved.

LEMMA 3. *Let* $A \overset{*}{\Rightarrow} xy_iz$, *where* A *is a nonterminal,* $i \geqslant 0$ *and either* $x = 0^\nu$, $z = 1^\mu$ *or* $x = 0^\nu$, $z = 1^N 0^\mu$ *or* $x = 1^\nu 0^N$, $z = 1^\mu$, *with* ν, $\mu \leqslant N$ *being integers. Then* $\mathrm{Ind}(\tau) \geqslant i + 1$ *for any derivation* τ *of* xy_iz *from* A.

This yields immediately $\mathrm{Ind}(G') \geqslant n$ and thus $\mathrm{Ind}(L_n) \geqslant n$.

Hence, to complete the proof of the Theorem, it remains only to prove the Lemma. The proof will be by induction on i. The case $i = 0$ is trivial. Assume that Lemma holds for 0, 1,..., $i - 1$, and let

$$A = W_0, W_1, \dots, W_s = xy_iz \tag{4}$$

be a derivation of xy_iz from A of the minimal index and as short as possible.

Since $N > 3m^{n_0+1}$, (2) implies that there exists the smallest integer i_0 such that W_{i_0+1} contains two nonterminals. Thus, $W_{i_0} = u_{i_0}A_{i_0}v_{i_0}$, where $u_{i_0}v_{i_0} \in \{0, 1\}^*$ and A_{i_0} is a nonterminal.

We claim that

$$|u_{i_0}| \leqslant |x| + N + m^{n_0+1}, \qquad |v_{i_0}| \leqslant |z| + N + m^{n_0+1}. \tag{5}$$

Assume that (5) does not hold. Then there must exist the smallest integer $i_1 \leqslant i_0$ such that $W_{i_1} = u_{i_1}A_{i_1}v_{i_1}$, $u_{i_1}v_{i_1} \in \{0, 1\}^*$ and

$$\text{either} \quad |u_{i_1}| > |x| + N + m^{n_0+1} \quad \text{or} \quad |v_{i_1}| > |z| + N + m^{n_0+1}. \tag{6}$$

Let z_{i_1} be such that $u_{i_1}z_{i_1}v_{i_1} = xy_iz$. Since $y_i = 0^{2N}y_{i-1}1^N0^Ny_{i-1}1^{2N}$ and $N > 3m^{n_0+1}$, from (6) it follows that $\psi(z_{i_1}) > m^{n_0}$ what contradicts to (2). Thus (5) holds.

Using (5), we can now complete the proof of Lemma. Clearly, $W_{i_0+1} = u_{i_0}u'BCv'v_{i_0}$ for some terminal words u' and v'. Since (4) was the shortest derivation of z from A among those with minimal index, none of the words $\bar{u}$, $\bar{v}$ where $B \overset{*}{\Rightarrow} \bar{u}$, $C \overset{*}{\Rightarrow} \bar{v}$, $z = u_{i_0}u'\bar{u}\bar{v}v'v_{i_0}$ is in $\{1\}^*\{0\}^*$. But it means that

$\bar{u}$ and $\bar{v}$ are of the form $x_1 y_{i-1} z_1$ with either $x_1 = 0^{\nu_1}$, $z_1 = 1^{\mu_1}$ or $x_1 = 0^{\nu_1}$, $z_1 = 1^N 0^{\mu_1}$ or $x_1 = 1^{\nu_1} 0^N$, $z_1 = 1^{\mu_1}$ and $\nu_1, \mu_1 \leqslant N$. By induction hypothesis $\text{Ind}(\tau_1) \geqslant i$, $\text{Ind}(\tau_2) \geqslant i$ for any derivations τ_1 of $\bar{u}$ from B and τ_2 of $\bar{v}$ from C. Hence $\text{Ind}(\tau) \geqslant i + 1$, completing the proof.

4. Undecidability

For any $n \in I \cup \{\infty\}$ there exist a context-free grammar G_n and a context-free language L_n such that $\text{Ind}(G_n) = n = \text{Ind}(L_n)$. For n infinite it follows from Salomaa's result (1969). For n finite, the existence of L_n was proved in the previous section and as G_n we can take the grammar with the rules $\sigma \to A^n$, $A \to a$.

On the other hand, as it will be shown in this section, for any $k \in I \cup \{\infty\}$ it is undecidable for a context-free grammar G whether or not $\text{Ind}(G) = k$ or whether or not $\text{Ind}(L(G)) = k$.

Two more results are proved in this section. If we think of Ind as being a criterion of complexity of grammars and languages, then they may be interpreted as follows: (i) It is undecidable whether a given grammar G is a simplest grammar for $L(G)$; (ii) There is no effective way to find a simplest grammar for $L(G)$, given a grammar G.

All these results will follow easily from the following lemma.

To simplify the ensuing discussion, let us denote by $P(x, y)$ the predicate which holds true if and only if x and y are n-tuples of nonempty words for some integer n and the post-correspondence problem for x and y has a solution.

Lemma 4. *For any n-tuples $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ of nonempty words over $\Sigma = \{0, 1\}$, and context-free grammars G_1 and G_2 with $L_i = L(G_i)$, a grammar G can be effectively found such that $L(G) = L' \cup L''$, where*

$$L' = \{ucwcv^R;\ u, v \in \Sigma,\ u \neq v,\ w \in L_1\}$$

and

$$L'' = \{x_{i_1} \cdots x_{i_k} cwc y_{i_k}^R \cdots y_{i_1}^R;\ k \geqslant 1,\ w \in L_2\}^2.$$

Furthermore, if $L_2 \subseteq L_1$, then

$$\text{Ind}(G) \begin{array}{l} \leqslant \text{Ind}(G_1) \text{ if } P(x, y) \text{ does not hold} \\ = \max\{\text{Ind}(G_1), \text{Ind}(G_2)\} \text{ if } P(x, y) \text{ holds,} \end{array}$$

[2] For a word $x = x_1 x_2 \cdots x_n$, $x^R = x_n \cdots x_2 x_1$.

and if, moreover, $\mathrm{Ind}(L_1) \leqslant \mathrm{Ind}(L_2) = \mathrm{Ind}(G_2)$, *then*

$$\mathrm{Ind}(L(G)) \begin{array}{l} \leqslant \mathrm{Ind}(L_1) \text{ if } P(x, y) \text{ does not hold} \\ = \mathrm{Ind}(L_2) \text{ if } P(x, y) \text{ holds.} \end{array}$$

Proof. Let $G_i = \langle V_i, \Sigma, P_i, S_i \rangle$, $V_1 \cap V_2 = \theta$, $A_1, A_2, A_3, S', S'', S \notin V_1 \cup V_2$.

Let G have the initial symbol S and productions $P_1 \cup P_2$ together with

$$\begin{aligned} &S \to S' \mid S'', \\ &S' \to 0S'0 \mid 1S'1 \mid 0A_1 \mid 1A_1 \mid 0A_21 \mid 1A_20 \mid A_30 \mid A_31, \\ &A_1 \to 0A_1 \mid 1A_1 \mid cS_1c, \\ &A_2 \to A_20 \mid A_21 \mid A_1, \\ &A_3 \to A_30 \mid A_31 \mid cS_1c, \\ &S'' \to x_iS''y_i^R \mid x_icS_2cy_i^R, \qquad 1 \leqslant i \leqslant n. \end{aligned}$$

Clearly, S' generates L' and S'' generates L''.

Let now $L_2 \subseteq L_1$. If $P(x, y)$ does not hold, then $L'' \subseteq L'$ and, therefore, $\mathrm{Ind}(G) \leqslant \mathrm{Ind}(G_1)$. If $P(x, y)$ holds, then $\mathrm{Ind}(G) = \max\{\mathrm{Ind}(G_1), \mathrm{Ind}(G_2)\}$.

To prove the last assertion of the Lemma, we will use the fact that $\mathrm{Ind}(L \cap R) \leqslant \mathrm{Ind}(L)$ if L is a CFL and R a regular set. It can be shown easily going through a standard proof of the theorem that the intersection of a language L and a regular set R is again a language (see, for example, Ginsburg, 1966).

Let now $\mathrm{Ind}(L_1) \leqslant \mathrm{Ind}(L_2) = \mathrm{Ind}(G_2)$. If $P(x, y)$ does not hold, then $\mathrm{Ind}(L(G)) \leqslant \mathrm{Ind}(L_1)$.

If $P(x, y)$ holds, then there are indices $i_1, \ldots, i_k$ such that

$$x_{i_1}x_{i_2} \cdots x_{i_k} = y_{i_1}y_{i_2} \cdots y_{i_k}.$$

Consider now the regular set

$$R = x_{i_1}x_{i_2} \cdots x_{i_k}c\{0, 1\}^*cy_{i_k}^R \cdots y_{i_2}^R y_{i_1}^R.$$

The intersection of R and $L(G)$ has the form

$$L(G) \cap R = x_{i_1} \cdots x_{i_k}cL_2cy_{i_k}^R \cdots y_{i_1}^R.$$

One can easily prove that if L is a language and a a symbol, then $\mathrm{Ind}(L) = \mathrm{Ind}(aL) = \mathrm{Ind}(La)$. Thus $\mathrm{Ind}(L(G) \cap R) = \mathrm{Ind}(L_2)$ and we have $\mathrm{Ind}(L(G)) \geqslant \mathrm{Ind}(L_2)$. On the other hand, from the construction of G it follows immediately that $\text{Ind}(L(G)) \leqslant \text{Ind}(G) \leqslant \text{Ind}(G_2) = \text{Ind}(L_2)$. Thus $\text{Ind}(L(G)) = \text{Ind}(L_2)$ and this completes the proof of the Lemma.

THEOREM 5. *Let $n \in I \cup \{\infty\}$. It is undecidable for an arbitrary grammar G whether or not* $\text{Ind}(G) = n$.

THEOREM 6. *Let $n \in I \cup \{\infty\}$. It is undecidable for an arbitrary grammar G whether or not* $\text{Ind}(L(G)) = n$.

For $n > 1$, the theorems follow from the Lemma 4 by taking G_1 to be the grammar $S_1 \to S_1 0$, $S_1 \to S_1 1$, $S_1 \to \epsilon$ and G_2 to be a grammar with $\text{Ind}(G_2) = \text{Ind}(L(G_2)) = n$. As a byproduct we get the theorems for $n = 1$.

COROLLARY 7. *There is no effective way to determine* $\text{Ind}(L(G))$ *for an arbitrary grammar G.*

THEOREM 8. *It is undecidable for an arbitrary grammar G whether or not* $\text{Ind}(G) = \text{Ind}(L(G))$.

Proof. Let us take as G_1 the grammar with the rules $S_1 \to S_1 S_1$, $S_1 \to 0$, $S_1 \to 1$, $S_1 \to \epsilon$ and as G_2 a grammar such that $\text{Ind}(G_2) = \text{Ind}(L(G_2)) = \infty$. Now $\text{Ind}(G) = \text{Ind}(L(G))$ for the grammar G from the Lemma 4 if and only if $P(x, y)$ holds. Hence the Theorem.

By using the same construction as in the proof of foregoing theorem, we get

COROLLARY 9. *There is no effective way to construct for an arbitrary grammar G, a grammar G' such that $L(G) = L(G')$ and* $\text{Ind}(G') = \text{Ind}(L(G))$.

5. MODIFICATION

The index of a grammar G represents the maximal number of nonterminals which may occur simultaneously in derivation steps in the derivations of elements in $L(G)$. However, there is no restriction as to how the nonterminals are spread out in words. In this section, we shall try to put some restriction on the distance between two nonterminals in a derivation step.

By $\text{Ind}'(G)$ we will mean the smallest integer k (if such a k does exist; otherwise we put $\text{Ind}'(G) = \infty$) such that for every $x \in L(G)$ there is a derivation $\sigma = w_0, w_1, \ldots, w_k = x$ in G such that each $w_i = u_i \alpha_i v_i$, where $u_i v_i$ is a terminal word and $|\alpha_i| \leqslant k$. Let $\text{Ind}'(L) = \min\{\text{Ind}'(G); L(G) = L\}$ for a language L.

Clearly, for every $n \in I \cup \{\infty\}$ there is a grammar G_n such that $\mathrm{Ind}'(G_n) = n$. Using the technique of the proof of Theorem 5, one can show that if $n \in I \cup \{\infty\}$, then it is undecidable for an arbitrary grammar G whether $\mathrm{Ind}'(G) = n$. Since $\mathrm{Ind}(L) \leqslant \mathrm{Ind}'(L)$ for every language L, $\mathrm{Ind}'(L_0) = \infty$. However, for every language L either $\mathrm{Ind}'(L) = \infty$ or $\mathrm{Ind}'(L) = 1$, and, therefore, the criterion Ind′ does not induce an infinite hierarchy of context-free languages. In order to show that $\mathrm{Ind}'(L) < \infty$ implies $\mathrm{Ind}'(L) = 1$ one can proceed as follows. Let G be a grammar such that $\mathrm{Ind}'(G) = k < \infty$ and $G = \langle V, \Sigma, P, \sigma\rangle$. Let us form a new grammar $G' = \langle V', \Sigma, P', \sigma\rangle$ by taking as new nonterminals the symbols $[\alpha]$, where $\alpha \in V^*$, $|\alpha| \leqslant k$ and the first and the last symbol of α are nonterminals. If, in G, $\alpha \Rightarrow u\beta v$, $|\beta| \leqslant k$, $uv \in \Sigma^*$, β starts and ends with nonterminal symbols, then we put to P' the production $[\alpha] \to u[\beta]v$ and all productions of P' are formed in that way. Clearly, $L(G') = L(G)$, $\mathrm{Ind}'(G') = 1$. Therefore, $\mathrm{Ind}'(L) < \infty$ if and only if L is a linear language. From that and from Greibach (1966) it follows that it is undecidable for an arbitrary grammar G whether $\mathrm{Ind}'(L(G)) = \infty$ (or $\mathrm{Ind}'(L(G)) = 1$).

Added in proof: Salomaa's problem was solved also by N. D. Jones in *Information and Control* **16**, 201–202.

Acknowledgment

The author wish to thank the referee for his suggestions concerning the exposition of Section 4.

Received: June 15, 1970; Revised: March 8, 1971

References

Brainerd, B. (1968), An analog of a theorem about context-free languages, *Information and Control* **11**, 561–567.

Ginsburg, S. (1966), The mathematical theory of context-free languages, McGraw-Hill, New York.

Ginsburg, S., and Spanier, E. H. (1968), Derivation-bounded Languages, *J. Comput. System Sci.* **2**, 228–250.

Greibach, S. A. (1966), The unsolvability of the recognition of linear context-free languages, *J. Ass. Comput. Mach.* **12**, 42–52.

Greibach, S. A. (1969), An infinite hierarchy of context-free languages, *JACM* **16**, 91–106.

Gruska, J. (1969), Some classifications of context-free languages, *Information and Control* **14**, 152–173.

Salomaa, A. (1969), On the index of context-free grammars and languages, *Information and Control* **14**, 474–477.